The Detection of Parkinson's Disease From Speech Using Voice Source Information

N. P. Narendra, Björn Schuller, *Fellow, IEEE*, and Paavo Alku, *Fellow, IEEE*

***Abstract*—Developing automatic methods to detect Parkinson's disease (PD) from speech has attracted increasing interest as these techniques can potentially be used in telemonitoring health applications. This article studies the utilization of voice source information in the detection of PD using two classifier architectures: traditional pipeline approach and end-to-end approach. The former consists of feature extraction and classifier stages. In feature extraction, the baseline acoustic features—consisting of articulation, phonation, and prosody features—were computed and voice source information was extracted using glottal features that were estimated by iterative adaptive inverse filtering (IAIF) and quasi-closed phase (QCP) glottal inverse filtering methods. Support vector machine classifiers were developed utilizing the baseline and glottal features extracted from every speech utterance and the corresponding *healthy/PD* labels. The end-to-end approach uses deep learning models which were trained using both raw speech waveforms and raw voice source waveforms. In the latter, two glottal inverse filtering methods (IAIF and QCP) and zero frequency filtering method were utilized. The deep learning architecture consists of a combination of convolutional layers followed by a multilayer perceptron. Experiments were performed using PC-GITA speech database. From the traditional pipeline systems, the highest classification accuracy (67.93%) was given by combination of baseline and QCP-based glottal features. From the end-to-end-systems, the highest accuracy (68.56%) was given by the system trained using QCP-based glottal flow signals. Even though classification accuracies were modest for all systems, the study is encouraging as the extraction of voice source information was found to be most effective in both approaches.**

***Index Terms*—Parkinson's disease, glottal features, glottal source estimation, support vector machines, end-to-end systems.**

I. Introduction

PARKINSON'S disease (PD) is the second most common neurodegenerative disorder after Alzheimer's disease [1]. PD primarily affects dopaminergic neurons in the substantia nigra of the brain, causing the loss of the neurotransmitter dopamine [2]. PD is diagnosed based on the occurrence of four gross motor dysfunctions, which include bradykinesia, rigidity, resting tremor, and postural instability [3]. However, by the time these dysfunctions are manifested in clinical diagnosis, up to 50% of the dopaminergic neurons are damaged beyond recovery, and the damage of the neurons has been found to increase rapidly during the four year period after diagnosis [4]. Any neuroprotective therapy performed after clinical diagnosis will be too late to effectively slow down neurodegeneration. Therefore, the early detection of PD in its prodromal stages is essential. Among many other symptoms, speech disorders are manifested in PD patients at the prodromal stages as early as five years before the occurrence of gross motor dysfunctions [5]. Speech disorders caused by PD can be characterized by symptoms such as reduced vocal tract volume and tongue flexibility, inappropriate pauses, impairments in voice quality, and reduction in pitch range and voice intensity [6]. Speech-based assessment of PD has attracted increasing interest among researchers as an automatic, low-cost, and easy-to-administer method for detecting early PD. The speech-based assessment of PD involves two main tasks: the detection of people with Parkinson's disease (PWP) from healthy speakers (binary classification) [7], [8] and the classification of the severity of PD (which consists of both multi-class classification and regression problems) [9], [10]. Of the two tasks above, this study focuses on the first one: the detection of PD by classifying speech signals into those produced by PWP and those produced by healthy speakers.

The detection of PD from speech can be used as an objective tool in non-invasive diagnosis, and therefore, the automatic detection of PD from speech is an important research topic. The detection of PD performed in an unobtrusive way has the potential of enhancing healthcare dramatically. The main advantage of non-invasive measurements is that the diagnosis can be done away from the hospital, which reduces the inconvenience and cost of the physical visits of PD patients for medical examination [11]. This advantage makes PD detection from speech one of the most preferable candidates for applications involving on-time screening and remote health monitoring [7], [12]. In the related literature, PD detection tools (such as Apkinson [13], the Johns Hopkins system [14], and mPower [15]), which are readily implementable in smart phones, have been proposed. In order to build health monitoring systems like these, different methods to automatically detect PD from speech have been developed. The PD detection methods proposed in the literature can mainly be divided into two approaches: *traditional pipeline* systems and *end-to-end* systems. In traditional pipeline systems, hand-crafted features that are obtained from speech are utilized

Manuscript received September 3, 2020; revised February 26, 2021; accepted April 29, 2021. Date of publication May 7, 2021; date of current version June 14, 2021. This work was supported by the Academy of Finland under Grant 330139. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zheng-Hua Tan PhD. *(Corresponding author: N.P. Narendra.)*

N. P. Narendra and Paavo Alku are with the Department of Signal Processing, and Acoustics, Aalto University, 00076 Espoo, Finland (e-mail: narendrasince1987@gmail.com; paavo.alku@aalto.fi).

Björn Schuller is with the GLAM (Group on Language, Audio and Music), Imperial College London, SW7 2AZ London, U.K, and also with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany (e-mail: schuller@ieee.org).

Digital Object Identifier 10.1109/TASLP.2021.3078364

to train machine learning methods to predict one of the two labels ($PD/healthy$). In end-to-end systems, the use of hand-crafted features is replaced by training deep learning models to directly map the raw speech signal (or the spectrogram) to the output labels ($PD/healthy$). The existing literature of these two approaches in the detection of PD is briefly reviewed below.

In using the traditional pipeline approach, the existing studies typically first select features that aim to characterize the impairments of speech signals of PWP. The majority of the works in the detection of PD characterizes speech impairments in terms of three main factors: articulation, phonation, and prosody [12], [16], [17]. In order to assess articulation impairments, features related to the vowel space area, vowel articulation index, the formant centralization ratio, and onset energy, as well as formant and spectral energy, have been used [17]–[21]. The assessment of impairments in phonation has been studied using features measuring perturbation such as the shimmer, the jitter, the amplitude perturbation quotient, the pitch perturbation quotient, the harmonic-to-noise ratio (HNR), and non-linear dynamics measures from the speech signal [7], [8], [19], [22]. Prosody impairments have been assessed using features based on pitch and energy contours, rhythm patterns, and duration [23]–[25]. In [26], a large set of acoustic features called ComParE [27] has been used to provide challenge baselines for the estimation of the neurological state of patients with PD. In addition to the features capturing articulation, phonation and prosody, features representing the mode of vibration of the vocal folds are also crucial as abnormalities in vocal fold closure patterns have been observed in PWP through laryngeal videoscopic examinations [28]. In order to capture irregularities in the airflow excitation signal generated by the vocal folds, the glottal flow, time-domain, and frequency-domain features have been extracted from the glottal flow [29], [30]. For estimating the glottal source from speech, glottal inverse filtering (GIF) must be used [31]. The existing studies on PD detection [29], [30] have used iterative adaptive inverse filtering (IAIF) [32] as the GIF method. Apart from features extracted from speech signals, the machine learning algorithm used as the classifier is also important. The majority of the studies in detection of PD using the traditional pipeline approach have used a support vector machine (SVM) as a classifier [7], [8], [18], [21], [33]. However, a few works have also utilized other machine learning algorithms such as random forest [16] and decision trees [22]. Even though the existing studies have investigated a large number of different features characterizing impairments in the speech signals of PWP, we argue that there is still a need for effective and robust features that are capable of distinguishing the speech of PWP from that of healthy controls.

The use of traditional pipeline systems in the detection of PD is justified because the approach is model-driven and based on features whose utilisation calls for understanding how the disease affects speech production (e.g. vocal folds, vocal tract, prosody etc.). Therefore, the features can be used not only by automatic, computerised detection systems but also, for example, by clinicians to get valuable knowledge about the speech production mechanism to understand which particular functions of the mechanism have been affected by the disease. The importance of this knowledge extraction embedded in the classical pipeline approach should not be underestimated despite that the current trend in the study area, the use of end-to-end systems (which will be described in the next paragraph) deliberately avoids the use of hand-crafted features. The main challenge in using the traditional, model-based pipeline approach with a separate feature extraction part is that the features need to be selected prior to their use in the model training and suitable, robust methods need to be designed to extract the selected features.

As an alternative to the traditional pipeline approach, end-to-end systems have recently been successfully utilized in the detection of PD [34], [35]. The end-to-end systems used in recent studies utilize speech spectra obtained from offset and onset transition regions for training deep learning models [12], [34]–[36]. For developing deep learning models, previous studies considered a framework of convolutional neural networks (CNNs) and multilayer perceptrons (MLPs) [12], [34]–[36]. The deep learning models trained with speech spectra obtained from offset and onset transition regions have been observed to perform better than SVM classifiers developed with separate sets of articulation and prosody features [12], [34]. However, SVM classifiers developed with the combination of phonation, articulation, and prosody features have resulted in better classification accuracy compared to deep learning models [12]. The lower performance of deep learning models has been attributed to smaller data sizes according to [12]. Apart from the speech spectrum, there are also other representations of speech that can be used to develop efficient deep learning models. Moreover, an organized comparison between end-to-end systems developed with different signal representations and traditional pipeline systems developed with different features sets is needed in the study area.

In contrast to classical pipeline systems, which are model-driven, end-to-end systems are in principle completely data-driven and they do not require any domain expertise in voice pathologies. However, end-to-end systems require large amounts of data to properly train deep learning models. Moreover, despite the fact that end-to-end systems may show excellent detection accuracy, it is difficult for the user (e.g. the clinician) to gain knowledge about the underlying reasons why a certain detection decision was made by the network.

The (estimated) glottal flow waveform, which carries voice source information, can potentially be utilized as a time-domain input signal in end-to-end systems, as demonstrated in previous works in both text-to-speech (TTS) synthesis [37] and pathological voice detection [38]. There are two justifications for studying the use of glottal flow signals in end-to-end systems. First, the glottal flow signal contains important information about the human speech production process that is related to voice quality [39], emotions [40], pathologies [41], and paralinguistics [42]. Second, compared to the time- and frequency-domain representations of the speech (pressure) signal, the glottal flow is a more elementary signal due to the absence of vocal tract resonances. With the utilization of such an elementary time-domain signal, end-to-end systems can be efficiently trained utilizing smaller amounts of training data, as indicated in [37]. This property of the glottal flow waveform is

particularly valuable when training deep learning networks in the area of speech-based health applications (such as the detection of PD from speech signals) because long recordings of speech training data from patients cannot be conducted as easily as they can from healthy speakers.

In the current study, the detection of PD from speech is studied by utilizing information carried by the voice source. The study has the following two main goals. The utilization of voice source information was found to improve the detection of PD in the recent study by Novotný et al. [30]. In their study, an old GIF method, IAIF [32], was used for the estimation of the glottal flow from the speech signal. Since our recent studies [43]–[45] indicate that a new GIF method—quasi-closed phase (QCP) analysis—gives more accurate estimates of the voice source, the $first$ goal of the present study is to compare how the underlying GIF method (i.e., IAIF vs. QCP analysis), which is used to compute the glottal flow waveforms to train the classifiers, affects the detection of PD. In [46], QCP analysis showed better accuracy in the estimation of the glottal flow in comparison to IAIF based on evaluations carried out using two types of synthetic vowels (produced by the Liljencrants-Fant model and by physical modeling of human voice production). Based on these previous results, we hypothesize that using QCP analysis in the traditional pipeline system instead of IAIF should result in better classification accuracy in the detection of PD. Since the previous studies [29], [30] investigating the use of voice source information in the detection of PD have exclusively used the traditional pipeline system, the $second$ goal of the study is to investigate the use of voice source waveforms in the end-to-end approach to the detection of PD. The second goal is justified by our recent study on the detection of dysarthria [38], which showed that using glottal flow waveforms instead of speech pressure waveforms as time-domain input signals for deep learning networks is able to improve the detection performance. In the second goal, our hypothesis is that the training of end-to-end systems using glottal flow waveforms should also result in better detection performance in the current task of detecting PD from speech.

This paper studies the use of voice source information in two approaches to PD detection (the traditional pipeline system and the modern end-to-end system). In the traditional pipeline approach, widely used acoustic features, consisting of articulation, phonation, and prosody features, are regarded as baseline features. Glottal features are extracted from the source waveforms obtained using two GIF methods: IAIF and QCP analysis. Using both the baseline and glottal features, separate sets of SVM classifiers are trained to output labels indicating $PD/healthy$. A comparison of the performance of the PD detection systems obtained using the two GIF methods is carried out. In the end-to-end approach, two kinds of time-domain waveforms are utilized for developing deep learning models. These two waveforms types include raw speech waveforms and raw voice source waveforms. In the latter, two GIF methods (IAIF and QCP analysis) based on the source-filter modeling of human speech production are used together with a third voice source estimatation method, zero frequency filtering (ZFF) [47], which is computed without explicitly computing the source-filter

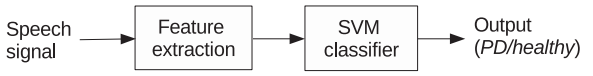

Fig. 1. The PD detection system based on the traditional pipeline approach.

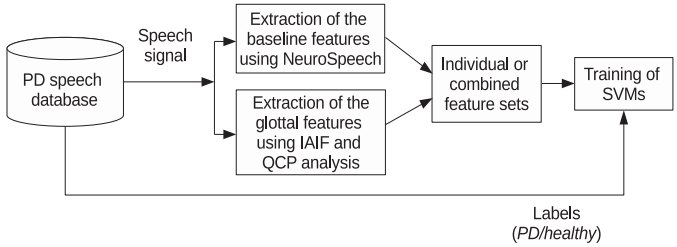

Fig. 2. The training stage of the PD detection system based on the traditional pipeline approach.

separation. The deep learning architecture used in this study consists of a combination of a CNN and an MLP. The performances of deep learning models developed using different time-domain waveforms are compared. In order to develop PD detection systems, a widely used PD speech database, PC-GITA [48], is utilized.

The remaining part of this paper is arranged as follows. The classical pipeline system and the end-to-end system studied in the current article are described in Sections II and III, respectively. Section IV provides details about the speech corpus and the experimental setup, and the evaluation results are reported in Section V. The results obtained are discussed in Section VI. Conclusions about the proposed method and some of the possible future works are provided in Section VII.

II. The PD Detection System Based on the Traditional Pipeline Approach

A. System Structure

To classify speech utterances into signals produced by PWP and into signals produced by healthy controls using the traditional pipeline approach, the PD detection system shown in Fig. 1 was developed. The system contains two major portions: feature extraction and an SVM classifier. In the feature extraction stage, selected features are extracted from the input speech signal. The features used in this work include articulation, phonation, and prosody features (collectively referred to as the $baseline\ features$) and glottal features. The baseline features are obtained from the speech signal using the NeuroSpeech toolkit [17]. The glottal features are computed from voice source waveforms, estimated using two GIF methods: IAIF [32] and QCP analysis [46]. Using both the baseline and glottal features, SVM classifiers are trained to predict one of the two labels ($PD/healthy$).

Fig. 2 shows the training phase of the PD detection system. First, a PD speech database (detailed in Section IV-A) consisting of utterances recorded from multiple speakers, both PWP and healthy people, is considered. From every utterance of the database, the baseline features are computed using the NeuroSpeech toolkit, as will be described in Section II-B. The glottal source waveforms are estimated from the speech signal

using the IAIF and QCP methods, as will be described in Section II-C1. From the estimated source waveforms, time- and frequency-domain glottal features are obtained as explained in Section II-C2. Using the pairs of extracted features and its corresponding output labels ($PD/healthy$), SVM classifiers are trained. Considering both the baseline and glottal feature sets, as well as their combinations, separate SVM classifiers are trained.

After training, the SVM classifiers are used to detect the presence of PD from speech signals. In order to test the developed PD detection systems, the same set of features that were used to train the classifiers are extracted from test utterances. The extracted features are used as inputs to the SVM classifiers that finally output the $PD/healthy$ labels.

B. Baseline Features

The baseline features used in this study consist of features characterizing different aspects of articulation, phonation, and prosody. The main reason for using features characterizing these three issues is that the same feature extraction approach has been widely used in PD detection tasks [49], [50], and the corresponding features have been used as baselines in comparing various detection systems [12], [34]. The baseline features used in this work are primarily based on the studies by Orozco-Arroyave et al. [17], [22], [51]. The articulation, phonation, and prosody features used in the current study are briefly described below:

- **Articulation** is parameterized using Bark band energies (BBEs), Mel-frequency cepstral coefficients (MFCCs), the first derivative of the MFCCs, the second derivative of the MFCCs at onset and offset transitions, and the first and second formant. Altogether, 122 articulation parameters are extracted and they are processed with four statistical functionals (mean, standard deviation, kurtosis, and skewness), resulting in a set of 488 $articulation\ features$ per utterance.
- **Phonation** is parameterized using the first and second derivatives of the fundamental frequency, jitter, pitch perturbation quotient, shimmer, amplitude perturbation quotient, and log energy computed from voiced segments. These seven phonation parameters are processed with four statistical functionals (mean, standard deviation, kurtosis, and skewness), resulting in a set of 28 $phonation\ features$ per utterance.
- **Prosody** is parameterized using duration, fundamental frequency, and energy. Six statistical functionals (mean, standard deviation, minimum, maximum, kurtosis, and skewness) are applied to these prosody parameters resulting in a set of 103 $prosody\ features$ per utterance. Apart from four statistical functionals used in articulation and phonation parameterization, two additional functionals (minimum and maximum) are considered during the computation of prosody parameters as these functionals are observed to be effective in discriminating PD patients from healthy controls.

C. GIF Methods and Glottal Features

1) GIF Methods: In order to use voice source information in the traditional pipeline system developed, glottal features need to be computed by first estimating the glottal flow waveform from every speech utterance using a GIF method. In this study, two GIF methods (IAIF and QCP analysis), which will be described next, are used in the estimation of the voice source. It should be noted that the third method to be used to extract voice source information in the current study, ZFF, is not used with the detection system based on the traditional pipeline approach. The reason for this is that, as in the study by Novotny et al. [30], we only wanted to use those methods that aim to estimate the true glottal flow (i.e., GIF methods based on the source-filter separation of speech) in this part of the study.

IAIF [32] is an old GIF method that is based on first estimating the spectral tilt of the glottal source from speech using a simple two-stage approach and then estimating the vocal tract model with linear prediction (LP) using the signal from which the glottal tilt has been removed. The spectral tilt of the glottal source is computed over several glottal periods (including both glottal open and closed phases), and therefore the IAIF method does not call for the extraction of glottal closure instants (GCIs). During the past two decades, IAIF has been used in many areas, such as in parametric speech synthesis ([52], [53] [54], [55]), speaking style conversion [56], the detection of stress [57], and depression [58], as well as in emotion recognition [59], [60]. For a detailed description of the IAIF method, the reader is referred to Section II-B in the work of Raitio et al. [52].

QCP analysis [46] is a more recently developed GIF method compared to IAIF. QCP analysis was selected in the current study because it was determined as the best performing GIF method in a comparison to four known reference GIF methods in [46]. QCP analysis is based on closed phase (CP) analysis [61], which estimates the vocal tract model using speech samples during the CP of the glottal cycle. Instead of using a few CP samples in the computation of the vocal tract as in CP analysis, QCP analysis computes a more robust estimate of the vocal tract by taking advantage of all the samples of the input frame. As explained in [46], this is enabled by using weighted linear prediction (WLP) analysis in the computation of the vocal tract model by downgrading the effect of those samples during which the effect of the source is prominent. The downgrading is done by using a specific temporal weighting function, called the attenuated main excitation (AME) function [62], in WLP. The computation of the AME function calls for extracting GCIs. The AME function attenuates the speech samples corresponding to the (quasi)–open phase during the computation of WLP, which subsequently leads to better estimates of the vocal tract transfer function.

Both IAIF and QCP analysis are GIF methods, which have been used in many studies [29], [30], [52], [60], [63], [64]. Both of the methods estimate the glottal source by removing the effect of the vocal tract by inverse filtering the speech signal through a digital all-pole filter model of the vocal tract. In order to estimate the vocal tract model, the two GIF methods follow different methodologies as explained in the previous two paragraphs:

IAIF uses conventional LP as the all-pole modeling method while QCP analysis takes advantage of WLP. QCP analysis was shown in [46] to give more accurate estimates of the glottal source compared to IAIF and this improvement was due to better modeling of the vocal tract by WLP compared to conventional LP. Computational loads of both methods in the estimation of the vocal tract and in inverse filtering are small. However, unlike the IAIF method, QCP analysis calls for extracting GCIs to generate the AME function. This results in a slight increase in the computational load for QCP analysis but the additional overhead caused by estimating GCIs is acceptable given the improved accuracy of QCP analysis for different pitch ranges as shown in [46].

2) Glottal Features: The time- and frequency-domain characteristics of the glottal flow waveforms estimated by the two GIF methods described above are parameterized using a set of voice source parameters as follows. The glottal parameters considered in this study consist of 12 known voice source parameters that represent different properties of the glottal flow waveform [65], [66]. These glottal parameters have been used recently, for example, in dysarthric and dysphonic voice classification tasks [38], [43]. These parameters are briefly described below:

- **The amplitude quotient (AQ):** The AQ is the ratio of the peak-to-peak amplitude of the glottal flow and the minimum peak of the flow derivative.
- **The normalized amplitude quotient (NAQ):** The NAQ is obtained by dividing the AQ by the length of the glottal cycle.
- **The open quotient (OQ):** The OQ is computed as the relative portion of the open phase compared to the cycle duration. Two OQs, OQ1 and OQ2, are computed from primary and secondary glottal openings, respectively.
- **The Liljencrants-Fant open quotient (OQa):** The OQa is the OQ derived by matching the estimated flow derivative with the Liljencrants-Fant (LF) model.
- **The closing quotient (ClQ):** The ClQ is computed as the ratio of the duration of the closing phase to the total length of the period.
- **The quasi-open quotient (QOQ):** The QOQ is the amplitude-domain counterpart of the OQ parameter, where the open duration is calculated as the time during which the flow is above a set level, usually 50% above the minimum flow.
- **The speed quotient (SQ):** The SQ is the ratio of the duration of the opening phase to the duration of the closing phase. Two SQs, SQ1 and SQ2, are calculated from the primary and secondary glottal openings, respectively.
- **The difference between first two glottal harmonics (H1H2):** This is the difference between the amplitude values of the first and second harmonics, obtained from the spectrum of the glottal flow on the dB scale.
- **The parabolic spectral parameter (PSP):** The PSP is computed based on fitting a parabolic function to the low-frequency part of the pitch-synchronous spectrum of the glottal flow.
- **The harmonic richness factor (HRF):** The HRF is obtained as the ratio of the sum of the magnitudes of the glottal harmonics above the fundamental frequency to the magnitude of the glottal harmonic at the fundamental frequency on the dB scale.

The glottal parameters are extracted from all voiced frames of the input speech signal. The H1H2 and HRF are determined pitch-asynchronously for every frame and the remaining parameters are extracted for every glottal cycle and then averaged over the frame. The glottal parameters extracted from all frames of an utterance form a glottal parameter vector. From the glottal parameter vector and its delta, eight statistical measures are determined: minimum, maximum, range, mean, median, standard deviation, skewness, and kurtosis. This leads to a set of $(12 + 12) \times 8 = 192$ $glottal\ features$ per utterance.

In order to demonstrate the behaviour of the glottal parameters, a two-way analysis of variance (ANOVA) was computed using glottal parameters extracted from the speech of PWP and their healthy controls. In general, a two-way ANOVA analysis tests the null hypothesis, i.e., that there is no statistically significant difference in the means of two independent variables and that the effect of one independent variable does not depend on the effect of the other independent variable (i.e., there is no interaction effect). The speech was taken from a PD database, which will be described in Section IV-A. Two-way ANOVAs were computed by considering the individual glottal parameters as dependent variables, and considering the health state ($PD/healthy$) and GIF method (IAIF/QCP analysis) as independent variables. Each glottal parameter was averaged over the utterances of every speaker. Table I shows the ANOVA results. Most importantly, the table indicates that several of the glottal parameters (NAQ, AQ, ClQ, SQ1, and SQ2) show statistically significant differences ($p < 0.001$) between the PWP and their healthy controls. In addition, the last two columns of Table I show that F-statistic is close to zero and $p > 0.001$ for all the glottal parameters, indicating that the null hypotheses are validated and that there is no interaction effect between health state and GIF method. These statistical tests confirm that the glottal parameters contain voice source information that should help to distinguish the speech of PWP from the speech of healthy talkers. In addition, the results show that the choice of the GIF method affects the detection as well. Finally, Fig. 3 demonstrates the glottal parameters between PWP and their healthy controls by showing the scatter plots for four pairs of the glottal parameters obtained from 100 randomly selected utterances of these two speaker classes. The figure illustrates that the glottal parameters of the PWP are distributed more widely compared to the healthy controls. Even though there is an overlap in the parameter distributions between the two speaker classes, the figure demonstrates that the two classes can be distinguished even when the data is expressed using simple two-dimensional glottal parameter scatter plots.

Finally, we would like to point out that the extraction of glottal source information described in this section is computed only from voiced segments of speech, because these are the segments during which the vocal folds vibrate generating the glottal flow excitation. However, the baseline features (i.e., the articulation

TABLE I
THE RESULTS OF THE TWO-WAY ANOVAS. THE GLOTTAL FEATURES ARE CONSIDERED AS DEPENDENT VARIABLES, AND THE HEALTH STATE ($PD/healthy$) AND GIF METHOD (IAIF/QCP ANALYSIS) ARE CONSIDERED AS TWO INDEPENDENT VARIABLES. THE NUMBER OF DEGREES OF FREEDOM FOR BOTH INDEPENDENT VARIABLES IS 1

Glottal features	Health state ($PD/healthy$)		GIF method (IAIF/QCP analysis)		Interaction effect Health state * GIF method	
	F-statistic	p	F-statistic	p	F-statistic	p
NAQ	23.33	<0.001	221.69	<0.001	0.48	0.4885
H1H2	2	0.1587	0	0.9483	0	0.9462
PSP	2.99	0.0855	0.32	0.5709	0.07	0.7884
HRF	2.91	0.0899	0.01	0.938	0	0.9690
QOQ	3.41	0.0664	1.66	0.1989	0.15	0.6967
AQ	28.8	<0.001	93.35	<0.001	0.53	0.4655
ClQ	23.51	<0.001	112.1	<0.001	0	0.9652
SQ1	21.25	<0.001	57.77	<0.001	0.58	0.4471
SQ2	20.16	<0.001	38.2	<0.001	0.57	0.4527
OQ1	0.08	0.7727	69.3	<0.001	0.04	0.8477
OQ2	0.07	0.7907	31.92	<0.001	0.16	0.6864
OQa	0.86	0.3549	139.88	<0.001	0.68	0.4121

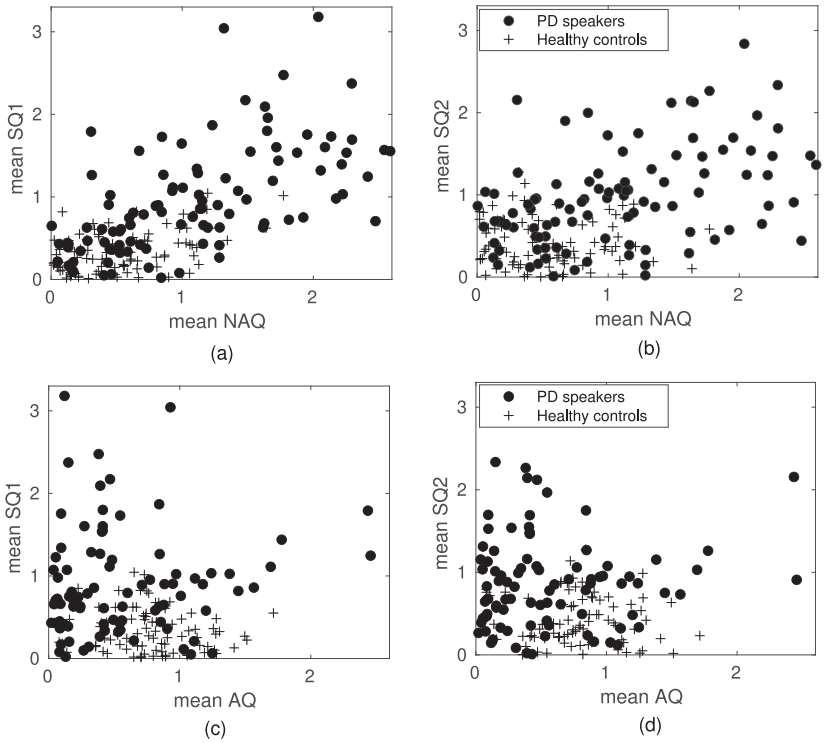

Fig. 3. The scatter plots of four pairs of glottal features (mean NAQ vs. mean SQ1, mean NAQ vs. mean SQ2, mean AQ vs. mean SQ1, and mean AQ vs. mean SQ2). The glottal features were extracted from the flow waveforms estimated using the QCP method.

and prosody features) are always extracted both from voiced and unvoiced segments.

III. THE PD DETECTION SYSTEM BASED ON THE END-TO-END APPROACH

The block diagram of the PD detection system developed using the end-to-end approach is shown in Fig. 4. The architecture of the end-to-end system contains several convolutional layers followed by an MLP. The end-to-end architecture followed in this work has been previously used in pathological voice classification and paralinguistic tasks [38], [67]–[69]. The end-to-end system takes a raw time-domain waveform (either the speech signal or the voice source waveform) as the input

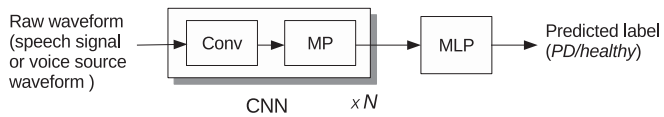

Fig. 4. The studied PD detection system based on the end-to-end approach. Conv: convolution, MP: Max pooling, MLP: multilinear perceptron.

and the system outputs the predicted label ($PD/healthy$). The input raw waveform enters into multiple convolutional layers. The CNN extracts suitable information from raw waveforms and this information is then passed through an MLP in order to estimate the label. The combination of the CNNs and MLP used in this study is jointly trained in a single framework.

In the current study, two types of raw time-domain waveforms, the speech (pressure) signal and the voice source waveform, are used to train end-to-end systems for the classification of PD. For the voice source, three different methods to compute the time-domain waveform are compared. Two of the three methods are based on GIF (IAIF and QCP analysis). The third one (ZFF [47]) is not a GIF method, but computes an approximate voice source waveform without explicitly using any source-filter model. ZFF takes advantage of the impulse-like nature of speech excitation. The method is based on the principle that discontinuity due to impulse excitation is reflected across all frequencies, including the zero frequency, and by designing a zero-resonance frequency filter, information about the discontinuities due to the impulse excitation can be obtained. The ZFF method has previously been used in the extraction of voice source information in different paralinguistic tasks such as in the detection of emotion and depression [69], [70]. More details about the ZFF method can be found in [47], [71], [72]. In using the end-to-end approach based on the raw speech signal, segments consisting of both voiced and unvoiced speech are used to train deep learning models.

Instead of using the entire utterance as the input of the deep learning models, the utterance is split into segments of constant duration that will be used as inputs to the systems. In the testing phase, the scores computed from each of the segments of the utterance are averaged and thresholded for obtaining the final binary decision ($PD/healthy$) for the utterance.

IV. Experiments

This section describes the experiments that were designed to study the use of voice source information in the detection of PD using the two detection systems described in Sections II and III. The evaluation was performed using the speech signals of the PC-GITA database. In using the traditional pipeline approach, SVM classifiers were developed using different combinations of the baseline and glottal features. The detection systems based on the end-to-end approach were developed using raw speech and voice source signals. The performance of the developed PD detection systems were evaluated using accuracy, sensitivity, and specificity measures. The brief descriptions related to the speech database and setups used in the experiments are provided separately in the next two sub-sections.

A. The PC-GITA Speech Database

The PC-GITA [48] speech database contains speech recordings in Spanish from 50 PD patients (25 male and 25 female) and 50 healthy control speakers (25 male and 25 female). The data was sampled at 44.1 kHz with a resolution of 16 bits. The PD patients have been diagnosed by neurologists. The healthy controls are free of any reported PD symptoms or other neurodegenerative disease. The speaker age varies from 31 years old to 86 years old. The recordings were performed in a sound-proof booth at Clínica Noel of Medellín in Colombia. The database includes speech collected using the following speaking tasks: (1) sustained phonation (vowels uttered in constant and modulated tones), (2) reading words aloud (25 isolated words taken from phonological inventory of Colombian Spanish), (3) diadochokinetic (DDK) exercises (repetition of the sequence of syllables: /pa-ta-ka/, /pe-ta-ka/, /pa-ka-ta/, /pa/, /ka/, /ta/) (4) reading sentences aloud (six simple and complex sentences), (5) reading a text (reading a dialogue between a doctor and a patient), and (6) giving a monologue (subjects are asked to speak about their daily routine and activities). In this study, speech tasks corresponding to continuous speech (i.e. DDK exercises, reading words and sentences aloud, and giving a monologue) are used. The speech signals were downsampled to 16 kHz in order to be used in the experiments of the present study.

B. The Experimental Setup

In order to train and evaluate the detection systems, a 10-fold cross-validation (CV) strategy was followed. In this strategy, the speech data was divided into 10 folds randomly. Of these 10 folds, 9 folds were used for training and the remaining one was used for testing. This process was repeated 10 times. Every speaker was used only once for testing and the same speaker was not used in both training and testing. The data partition followed in this work has been previously used in many PD detection studies [22], [51]. From the data used in training, 90% of every speaker's utterances was used to train the detection systems and the remaining 10% was used for validation.

For developing the traditional pipeline approach of PD detection systems, speech signals were processed in 30 ms frames with a 15 ms shift. The baseline and glottal features were computed from every utterance of the database. The baseline features consisting of articulation, phonation and prosody features were extracted using the NeuroSpeech toolkit [17] and the glottal features were extracted using the APARAT Toolbox [73]. In order to extract glottal features, the glottal source signals were first estimated by the GIF methods described in Section II.C (IAIF and QCP analysis). By referring to [52], the IAIF method was computed by using filter orders of $q = 10$ and $p = 24$ for the spectral tilt model of the glottal source and for the vocal tract model, respectively. All the LP analyses were computed in IAIF by using the autocorrelation method and the Hann window. QCP analysis estimates the glottal source by using WLP in vocal tract modeling and the computational steps reported in [46] were used in the current study. GCIs required for the computation of the AME function were estimated from speech using the GLOAT toolkit [74]. The order of the vocal tract model in QCP analysis was the same as in IAIF (i.e. 24).

The glottal features obtained using IAIF and QCP analysis are referred as "Glottal (IAIF)" and "Glottal (QCP)," respectively. From the baseline and glottal features, the global mean and global standard deviation were determined and these measures were used to normalize each of the individual features. The normalized features were used to train SVM classifiers. Distinct sets of SVM classifiers were trained using the baseline and glottal features, as well as their combination. SVM classifiers were trained with the Gaussian radial basis function kernel, and the kernel parameter γ and penalty parameter C were determined separately for every set of SVM classifier. The optimal values of C and γ were determined based on a grid search with their values

TABLE II
THE RANGES OF HYPER PARAMETERS FOR THE GRID SEARCH. NOTE THAT IN OUR NETWORK, ONLY ONE MLP LAYER IS USED

Parameters	Units	Range
Kernel size	samples	4 - 150
Number of filters	filters	4 - 100
Max Pooling kernel width	frames	2 - 10
Max Pooling kernel shift	frames	2 - 10
Number of units in MLP	units	10 - 200
batch size	examples	30 - 500
dropout		0 - 1
learning rate		0.0001-0.01
number of iterations		10 - 500

TABLE III
END-TO-END SYSTEM NETWORK ARCHITECTURE

Network	Configuration
CNN+MLP	conv1: filters = 16, kernel size = 64, Activation: Relu, Maxpooling: pool size = 2, strides = 2 conv2: filters = 32, kernel size = 32, Activation: Relu, Maxpooling: pool size = 2, strides = 2 conv3: filters = 64, kernel size = 16, Activation: Relu, Maxpooling: pool size = 2, strides = 2 MLP: 128 hidden units, activation: Relu fully connected output layer, activation: Sigmoid

varying from 10^{-3} to 10^{3} in multiples of 10 and the selection criteria being accuracy computed using the validation data.

For the end-to-end systems, the raw speech signals and the raw voice source waveforms (computed using IAIF, QCP analysis and ZFF) were processed in segments of 250 ms with a 50 ms shift. In order to compute ZFF, the method proposed in [47] was used. The speech signal was passed through a cascade of two ideal digital resonators located at 0 Hz, and then the trend in the resulting signal was subtracted in a window whose size equaled the average pitch period. The glottal flow waveforms obtained from IAIF and QCP analysis are denoted as "Glottal flow (IAIF)" and "Glottal flow (QCP)," respectively, and the source waveform obtained using ZFF is referred to as "Voice source (ZFF)". Both the raw speech signals and the raw voice source waveforms were split into segments of 250 ms. This duration was chosen based on similar experiments carried out in [38], [75], [76].

By utilizing four types of raw waveforms that were split into fixed length segments, CNN+MLP networks were trained separately. Before training the CNN+MLP networks, the hyper parameters of the network were chosen based on a grid search. The hyper parameters were varied over a certain range, and the values resulting in better accuracy were chosen. The hyper parameter ranges considered for the grid search are shown in Table II.

Table III provides the details about the network architecture. After every convolutional layer, a ReLu activation function and MaxPooling operation (pool size $=$ 2 and stride $=$ 2) were carried out. Batch normalization was used to make the training of the deep neural network faster and stable through the normalization of the input layer [77], and dropout was followed in every layer to avoid overfitting the deep neural networks [78]. In order to optimize the parameters of the end-to-end system, a stochastic gradient descent algorithm was utilized with the binary cross entropy error criterion. The total number of iterations used in training was 100, and an early stopping criterion was followed for five epochs with no improvement. The details about the parameters used in this work are provided in Table IV.

TABLE IV
THE PARAMETERS USED TO TRAIN THE END-TO-END SYSTEM

Parameters	Values
batch size	200
dropout	0.25
learning rate	0.01
number of iterations	100

TABLE V
DETECTION RESULTS COMPUTED FROM THE TRADITIONAL PIPELINE SYSTEMS DEVELOPED WITH THE BASELINE FEATURES (CONSISTING OF THE ARTICULATION, PHONATION, AND PROSODY FEATURES) AND GLOTTAL FEATURES

Feature set	Accuracy (%)	Sensitivity (%)	Specificity (%)
Articulation (art)	65.07	63.71	66.43
Phonation (phon)	62.71	60.43	65.00
Prosody (pro)	62.86	63.00	62.71
Baseline (art+phon+pro)	65.57	63.29	67.86
Glottal (IAIF)	63.64	57.43	69.86
Baseline + Glottal (IAIF)	67.00	64.71	69.29
Glottal (QCP)	64.64	58.86	**70.43**
Baseline + Glottal (QCP)	**67.93**	**69.71**	66.14

To assess the performance of the developed PD detection systems, three measures were considered: accuracy, specificity, and sensitivity. The accuracy was computed as the ratio between the number of speech utterances that are properly detected to the total number of utterances. Specificity was computed as the ratio between the number of correctly identified speech utterances spoken by the healthy speakers and the total number of utterances spoken by the healthy speakers. Sensitivity was determined as the ratio between the number of correctly classified speech utterances spoken by the PWP and the total number of utterances spoken by the PWP. For a good detection system, all three performance metrics should be high (ideally 100%). In addition to aiming at high values of sensitivity and specificity, indicating good performance in the identification of both PWP and their healthy controls, the difference between these two metrics should also be as low as possible for a good detection system, indicating that the system is not biased towards one of the two classes ($PF/healthy$). It should be noted that the data used for training and testing is balanced, that is, the data includes the same of number of PWP and healthy speakers and all the speakers produce an equal number of speech utterances.

V. RESULTS

Table V shows the detection results computed from the traditional pipeline systems developed with the different features. From the table, it can be noted that the detection results of the systems developed with the baseline features are slightly better compared with the glottal features obtained using both GIF methods (except in specificity for Glottal (IAIF) and Glottal (QCP)). Among the three baseline feature sets, the articulation features show the best detection result. The accuracies of the detection systems obtained with the Glottal (IAIF) and Glottal

TABLE VI
DETECTION RESULTS COMPUTED FROM THE END-TO-END SYSTEMS DEVELOPED WITH DIFFERENT RAW TIME-DOMAIN WAVEFORMS. THE THREE METRICS ARE AVERAGED OVER FIVE RUNS, AND THE MEAN AND STANDARD DEVIATION VALUES ARE PROVIDED

Input	Accuracy (%)	Sensitivity (%)	Specificity (%)
Speech signal	66.75 ± 2.10	58.06 ± 4.07	75.43 ± 1.85
Glottal flow (IAIF)	67.69 ± 1.03	61.00 ± 2.09	74.40 ± 2.05
Glottal flow (QCP)	**68.56 ± 0.87**	**63.40 ± 2.48**	73.73 ± 3.01
Voice source (ZFF)	66.28 ± 1.76	53.70 ± 4.39	**78.86** ± 1.83

(QCP) feature sets vary in the range of 63–64%. These accuracies indicate that the glottal source contains the relevant information required for the detection of PD. Among the glottal features obtained from the two GIF methods, the features obtained from QCP analysis show better detection results compared to those obtained from IAIF. Furthermore, by separately combining the two glottal features sets with the baseline features, the detection results are observed to be improved in both cases. This shows the complementary nature of the glottal features that results in better performance metrics (accuracy, specificity, and sensitivity) when combined with the baseline features. In a comparison of the two combinations of the baseline and glottal features, the Baseline + Glottal (QCP) feature set leads to the improved accuracy. By studying the sensitivity and specificity values of the classifiers developed with the different feature sets, it can be observed that in most of the cases, specificity values are moderately better compared with sensitivity values (except in the Prosody and Baseline + Glottal (QCP) feature sets). From Table V, it can be noted that by combining the glottal features with the baseline features both sensitivity and specificity improve and the difference between these two metrics is also reduced.

Table VI shows the detection results of the end-to-end systems developed with the different types of raw time-domain waveforms. The accuracies obtained using the raw glottal flow waveforms estimated by IAIF and QCP analysis are slightly better than those obtained using the raw speech signal and the ZFF-based voice source waveform. The system based on using the raw glottal flow estimated by QCP analysis shows modest improvements in accuracy and sensitivity compared to all other systems. Similarly to Table V, a comparison of the sensitivity and specificity values between the developed end-to-end systems shows that the specificity values are better than the sensitivity values. Moreover, Table VI indicates that despite the fact that all the systems show modest values in accuracy, sensitivity and specificity, the system trained using the QCP-based input waveform shows the lowest difference between sensitivity and specificity among all the systems.

Using the detection results reported in Tables V and VI, three evaluation metrics computed from the classifiers developed with the classical pipeline approach can be compared with those of the more modern end-to-end systems. With regard to accuracy and specificity, the best end-to-end system trained using the QCP-based glottal flow is observed to be moderately better (by an absolute improvement of about 1% in accuracy and 7% in specificity) than the best classical pipeline classifier trained using the Baseline + Glottal (QCP) features. However, the sensitivity of the best end-to-end system trained with the QCP-based glottal flow is lower (by an absolute decrease of about 6%) than that of the best classical pipeline system trained using the combination of the Baseline + Glottal (QCP) features. Among the classifiers developed with both the classical pipeline and end-to-end approaches, the features and flow waveforms obtained using the QCP method lead to moderately better results compared to both IAIF and ZFF.

VI. DISCUSSION

This paper studies the use of voice source information in the detection of PD by comparing the traditional, feature-based pipeline approach and the modern end-to-end approach. Both of the two approaches have their advantages and disadvantages. Since the traditional pipeline approach is based on representing speech signals with features, the approach benefits in principle from its capability to give better knowledge about how different parts of speech production (such as the vocal folds) contribute to the detection of PD. Clinicians and speech language pathologists can benefit from this knowledge in examining patients. In addition, the traditional pipeline system can be trained to produce good results with relatively small amounts of data [38]. The end-to-end approach does not call for using pre-defined sets of hand-crafted features and therefore it is an attractive choice in cases where computerised detection needs to be implemented by, for example, industrial players with no expertise in voice pathology. The end-to-end approach, however, requires large amounts of data for proper training of the deep learning models. It is also worth emphasising that while the amount of training data can be easily increased in areas such as in speech recognition and synthesis, where speech data is produced by healthy speakers, the same is not necessarily true in collecting pathological voice data because the data is recorded from patients whose health condition might be too weak to bear long recordings. Finally, though the end-to-end approach has shown better accuracy compared to the traditional pipeline approach in recent studies [12], [34], the end-to-end technology has been criticized [79] from a principal point of view because it provides a black box -type of solution with poor interpretability to the detection task. Suitability of particular approach is an open question and the reader can chose a particular approach based on his/her interest and field of application.

In this study, we focused on the effectiveness of voice source information in the detection of PD. Though the voice source has been used previously in detection tasks based on the traditional pipeline approach [29], [30], its usage in the end-to-end approach as studied in the current investigation has been not explored before. The results obtained both by the traditional systems and by the end-to-end systems did not, however, show high values of detection accuracy and the accuracy improvement obtained by using glottal source information can rather be characterised as modest. However, despite the modest improvement in the detection accuracy, the usage of glottal source information can be considered beneficial and promising according to the experiments conducted both for the traditional pipeline system and for the end-to-end system. In the case of the traditional

pipeline approach, the combination of the glottal features with the baseline features resulted in the best (yet modest) detection accuracy (67.93%). In the end-to-end system approach, deep learning models trained with glottal source signals were able to perform better compared to the models trained using the raw speech waveform. The best (but still modest) overall accuracy (68.56%) among all the systems studied was given by the end-to-end system that used the raw glottal flow computed by QCP analysis. The current study can be viewed as a reference point to carry out further research in Parkinson's disease detection using glottal source information.

Apart from the speech waveform, the usage of the glottal source signal as a raw waveform in end-to-end systems is justified because the voice pathologies affect the vocal folds (as reported in [41], [80]). The improved, yet modest accuracy obtained by using the raw glottal source in the current study demonstrates the importance of glottal source information in the end-to-end approach. In addition, the interesting result of the current study is that end-to-end system trained with glottal source signals was shown to perform better than the system trained using raw speech waveforms for a given available training data. The main reason for this is as follows. In the raw speech waveform, glottal information is also embedded because the speech waveform is a result of filtering the glottal excitation with the vocal tract. However, in addition to glottal information, raw speech waveforms also include phonemic and speaker-specific information that is brought about by the vocal tract. Involvement of phonemic and speaker-specific information makes learning of the detection problem more difficult for deep learning networks if there is only a relatively small amount of training data available (which is the case in the present study due to training the systems with speech of patients suffering from PD).

The PC-GITA database used in the current experiments contains Spanish speech utterances collected from different speaking tasks (DDK exercises, production of isolated words, production of sentences and reading a monologue). The DDK tasks used in this study include continuous repetitions of syllable sequences: /pa-ta-ka/, /pe-ta-ka/, /pa-ka-ta/, /pa/, /ka/, /ta/. For Spanish, which is a strongly syllable-rhythmic language, certain DDK tasks may not be well suited to differentiate PD from HC. Therefore, it might be that the current results cannot be fully generalised to languages which are less syllable-rhythmic. Similar kind of experimentation using DDK exercises considered in this study has, however, been followed in previous works in the detection of PD [25], [34]. More research should be devoted in the future to better understand the role of different speaking tasks (specifically related to different DDK exercises) in the automatic detection of PD.

VII. Conclusion

The automatic detection of PD from speech signals was investigated using voice source information and using two detection system architectures (the classical pipeline approach and the end-to-end approach). In the classical pipeline approach, SVM classifiers were developed to estimate $PD/healthy$ labels using known baseline features characterizing articulation, phonation, and prosody and using glottal features to describe voice source information. The baseline features were obtained from the NeuroSpeech toolkit and the glottal features were computed from the flow waveforms estimated using two GIF methods (IAIF and QCP analysis). The experiments indicated that using QCP analysis instead of IAIF as a GIF method improved the detection performance when the system was trained using voice source information alone. Most importantly, the study showed that the accuracy of the SVM-based detection system trained with the baseline features improved from 65% to 67% when voice source information was merged with the baseline features. In the end-to-end system approach, deep learning models based on combining CNNs and MLP were developed using raw speech and raw voice source waveforms obtained with two GIF methods (IAIF, QCP analysis), and with one method (ZFF), which yields the approximated voice source without using source-filter separation. The results showed that the detection performance was moderately better for the end-to-end systems developed with QCP-based glottal flow compared to IAIF and ZFF. In addition, the best overall accuracy (of 68%) among all the systems compared was achieved by using the end-to-end system that uses the QCP-computed raw glottal flow waveform as the input signal.

The present work studies the importance of voice source information in PD detection by comparing the traditional pipeline approach and the modern end-to-end approach. The work showed the effectiveness of the glottal features and glottal flow waveforms obtained using the QCP method in both of the two approaches. The present work can be extended, for example, as follows. In addition to the binary classification task studied in the current investigation, the method developed can be extended to predict the neurological state of PD patients. In addition to PD, the effectiveness of voice source information can be studied in other neurodegenerative diseases such as Alzheimer's disease and amyotrophic lateral sclerosis. Also, one can investigate the correlation of the traditional features and the convolutional layers' activations in order to investigate what the end-to-end approach has learnt to model.

References

[1] M. C. de Rijk *et al.*, "Prevalence of Parkinson's disease in Europe: A collaborative study of population-based cohorts. Neurologic diseases in the elderly research group," *Neurology*, vol. 54, no. 11, pp. S21–S23, 2000.

[2] W. Poewe *et al.*, "Parkinson disease," *Nature Rev. Dis. Primers*, vol. 23, no. 3, 2017, Art. no. 17013.

[3] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *J. Neural. Neurosurg. Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[4] M. C. Rodriguez-Oroz *et al.*, "Initial clinical manifestations of Parkinson's disease: Features and pathophysiological mechanisms," *Lancet Neurol.*, vol. 8, no. 12, pp. 1128–1139, 2009.

[5] B. Harela, M. Cannizzaro, and P. J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study," *Brain Cogn.*, vol. 56, no. 1, pp. 24–29, 2004.

[6] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients," *J. Speech Hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[7] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–1022, Apr. 2009.

[8] A. Tsanas, M. A. Little, P. E. McSharry, J. L. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, May 2012.

[9] G. An, D. G. Brizan, M. Ma, M. Morales, A. R. Syed, and A. Rosenberg, "Automatic recognition of unified Parkinson's disease rating from speech with acoustic, i-vector and phonotactic features," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 508–512.

[10] W. Caesarendra, F. T. Putri, M. Ariyanto, and J. D. Setiawan, "Pattern recognition methods for multi stage classification of Parkinson's disease utilizing voice features," in *Proc. IEEE Int. Conf. Adv. Intell. Mechatronics*, 2015, pp. 802–807.

[11] C. G. Goetz *et al.*, "Testing objective measures of motor impairment in early Parkinson's disease: Feasibility study of an at-home testing device," *Movement Disorders*, vol. 24, no. 4, pp. 551–556, 2009.

[12] T. Arias-Vergara, J. C. Vásquez-Correa, J. R. Orozco-Arroyave, P. Klumpp, and E. Nöth, "Unobtrusive monitoring of speech impairments of Parkinson's disease patients through mobile devices," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6004–6008.

[13] P. Klumpp, T. Janu, T. Arias-Vergara, J. C. V. Correa, J. R. Orozco-Arroyave, and E. Nöth, "Apkinson - A mobile monitoring solution for Parkinson's disease," in *Proc. INTERSPEECH*, 2017, pp. 1839–1843.

[14] S. Arora *et al.*, "Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study," *Parkinsonism Related Disorders*, vol. 21, no. 6, pp. 650–653, 2015.

[15] B. M. Bot *et al.*, "The mPower study, Parkinson disease mobile data collected using ResearchKit," *Sci. Data*, vol. 3, no. 160011, pp. 1–9, 2016.

[16] E. Vaiciukynas, A. Verikas, A. Gelzinis, and M. Bacauskiene, "Detecting Parkinson's disease from sustained phonation and speech signals," *PLoS One*, vol. 12, no. 10, pp. 1–16, 2017.

[17] J. R. Orozco-Arroyave *et al.*, "NeuroSpeech: An open-source software for Parkinson's speech analysis," *Digit. Signal Process.*, vol. 77, no. 1, pp. 207–221, 2018.

[18] M. Novotný, J. Rusz, R. Cmejla, and E. Ruzicka, "Automatic evaluation of articulatory disorders in Parkinson's disease," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 9, pp. 1366–1378, Sep. 2014.

[19] J. R. Orozco-Arroyave, *Analysis of Speech of People With Parkinson's Disease*, 1st ed. Berlin, Germany: Logos Verlag Berlin, 2016.

[20] B. Karan, S. S. Sahu, J. R. Orozco-Arroyave, and K. Mahto, "Hilbert spectrum analysis for automatic detection and evaluation of Parkinson's speech," *Biomed. Signal Process. Control*, vol. 61, no. 3, pp. 1–11, 2020.

[21] F. O. López-Pabón, T. Arias-Vergara, and J. R. Orozco-Arroyave, "Cepstral analysis and Hilbert-Huang transform for automatic detection of Parkinson's disease," *TecnoLogicas*, vol. 23, no. 47, pp. 93–108, 2020.

[22] D. Hemmerling, J. R. Orozco-Arroyave, A. Skalski, J. Gajda, and E. Nöth, "Automatic detection of Parkinson's disease based on modulated vowels," in *Proc. INTERSPEECH*, 2016, pp. 1190–1194.

[23] S. Skodda, W. Visser, and U. Schlegel, "Gender-related patterns of dysprosody in Parkinson disease and correlation between speech variables and motor symptoms," *J. Voice*, vol. 25, no. 1, pp. 76–82, 2011.

[24] L. Liu, M. Jian, and W. Gu, "Prosodic characteristics of mandarin declarative and interrogative utterances in Parkinson's disease," in *Proc. INTERSPEECH*, 2019, pp. 3870–3874.

[25] A. Rueda, J. C. Vásquez-Correa, C. D. Rios-Urrego, J. R. Orozco-Arroyave, S. Krishnan, and E. Noth, "Feature representation of pathophysiology of Parkinsonian dysarthria," in *Proc. INTERSPEECH*, 2019, pp. 3048–3052.

[26] B. Schuller *et al.*, "The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's and eating condition," in *Proc. INTERSPEECH*, 2015, pp. 478–482.

[27] B. Schuller *et al.*, "The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load," in *Proc. INTERSPEECH*, 2014, pp. 427–431.

[28] I. Midi, M. Dogan, M. Koseoglu, G. Can, M. A. Sehitoglu, and D. I. Gunal, "Voice abnormalities and their relation with motor dysfunction in Parkinson's disease," *Acta Neurologica Scandinavica*, vol. 117, no. 1, pp. 26–34, 2008.

[29] E. A. Belalcázar-Bolaños, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, T. Haderlein, and E. Noth, "Glottal flow patterns analyses for Parkinson's disease detection: Acoustic and nonlinear approaches," in *Proc. Int. Conf. Text, Speech, Dialogue*, 2016, pp. 400–407.

[30] M. Novotný, P. Dusek, I. Daly, E. Ruzicka, and J. Rusz, "Glottal source analysis of voice deficits in newly diagnosed drug-naive patients with Parkinson's disease: Correlation between acoustic speech characteristics and non-speech motor performance," *Biomed. Signal Process. Control*, vol. 57, no. 1, pp. 1–9, 2020.

[31] P. Alku, "Glottal inverse filtering analysis of human voice production – A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[32] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2–3, pp. 109–118, 1992.

[33] A. Rueda and S. Krishnan, "Augmenting dysphonia voice using fourier-based synchrosqueezing transform for a CNN classifier," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6415–6419.

[34] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Noth, "Convolutional neural network to model articulation impairments in patients with Parkinson's disease," in *Proc. INTERSPEECH*, 2017, pp. 314–318.

[35] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, and E. Noth, "A multitask learning approach to assess the dysarthria severity in patients with Parkinson's disease," in *Proc. INTERSPEECH*, 2018, pp. 456–460.

[36] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Noth, "Multimodal assessment of Parkinson's disease: A deep learning approach," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 4, pp. 1618–1630, Jul. 2019.

[37] L. Juvela, B. Bollepalli, V. Tsiaras, and P. Alku, "GlotNet – A raw waveform model for the glottal excitation in statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 6, pp. 1019–1030, Jun. 2019.

[38] N. P. Narendra and P. Alku, "Glottal source information for pathological voice detection," *IEEE Access*, vol. 8, pp. 67745–67755, 2020.

[39] J. Kreiman *et al.*, "Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation," *J. Acoust. Soc. Amer.*, vol. 132, no. 4, pp. 2625–2632, 2012.

[40] M. Airas and P. Alku, "Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalised amplitude quotient," *Phonetica*, vol. 63, no. 1, pp. 26–46, 2006.

[41] D. H. Tsuji, A. Hachiya, M. E. Dajer, C. C. Ishikawa, M. T. Takahashi, and A. N. Montagnoli, "Improvement of vocal pathologies diagnosis using high-speed videolaryngoscopy," *Int. Arch. Otorhinolaryngol.*, vol. 18, no. 3, pp. 294–302, 2014.

[42] E. Moore, M. A. Clements, J. W. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 96–107, Jan. 2008.

[43] N. P. Narendra and P. Alku, "Dysarthric speech classification from coded telephone speech using glottal features," *Speech Commun.*, vol. 110, no. 1, pp. 47–55, 2019.

[44] N. P. Narendra and P. Alku, "Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features," *Comput. Speech Lang.*, vol. 65, no. 1, pp. 1–14, 2020.

[45] N. P. Narendra and P. Alku, "Automatic intelligibility assessment of dysarthric speech using glottal parameters," *Speech Commun.*, vol. 123, no. 1, pp. 1–9, 2020.

[46] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 596–607, Mar. 2014.

[47] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[48] J. R. Orozco-Arroyave, J. D. Arias-Londono, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Noth, "New spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *Proc. Lang. Resour. Eval. Conf.*, 2014, pp. 342–347.

[49] J. V. E. López, J. R. Orozco-Arroyave, and G. Gosztolya, "Assessing Parkinson's disease from speech using fisher vectors," in *Proc. INTERSPEECH*, 2019, pp. 3063–3067.

[50] N. Garcia, J. R. Orozco-Arroyave, L. F. D'Haro, N. Dehak, and E. Noth, "Evaluation of the neurological state of people with Parkinson's disease using i-vectors," in *Proc. INTERSPEECH*, 2017, pp. 299–303.

[51] J. Orozco-Arroyave *et al.*, "Towards an automatic monitoring of the neurological state of Parkinson's patients from speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 6490–6495.

[52] T. Raitio *et al.*, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 153–165, Jan. 2011.
[53] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Comparing glottal flow-excited statistical parametric speech synthesis methods," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 7830–7834.
[54] T. Raitio *et al.*, "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *Proc. 22nd Eur. Signal Process. Conf.*, 2014, pp. 2290–2294.
[55] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5120–5124.
[56] O. Perrotin and I. V. McLoughlin, "Glottal flow synthesis for whisper-to-speech conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, no. 2, pp. 889–900, 2020.
[57] M. Sigmund, A. Prokes, and Z. Brabec, "Statistical analysis of glottal pulses in speech under psychological stress," in *Proc. Eur. Signal Process. Conf.*, 2008, pp. 1–5.
[58] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 574–586, Mar. 2011.
[59] V. Sethu, E. Ambikairajah, and J. Epps, "On the use of speech parameter contours for emotion recognition," *EURASIP J. Audio, Speech, Music Process.*, vol. 2013, no. 19, pp. 1–14, 2013.
[60] X. Yao, W. Bai, Y. Ren, X. Liu, and Z. Hui, "Exploration of glottal characteristics and the vocal folds behavior for the speech under emotion," *Neurocomputing*, vol. 410, no. 10, pp. 328–341, 2020.
[61] D. Wong, J. Markel, and A. Gray Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 4, pp. 350–355, Aug. 1979.
[62] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *J. Acoust. Soc. Amer.*, vol. 134, no. 2, pp. 1295–1313, 2013.
[63] P. Corcoran, A. Hensman, and B. Kirkpatrick, "Glottal flow analysis in parkinsonian speech," in *Proc. 12th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, 2019, pp. 116–123.
[64] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Comput. Speech Lang.*, vol. 28, no. 5, pp. 1117–1138, 2014.
[65] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Amer.*, vol. 90, no. 5, pp. 2394–2410, 1991.
[66] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parameterization of the glottal flow," *J. Acoust. Soc. Amer.*, vol. 112, no. 2, pp. 701–710, 2002.
[67] G. Trigeorgis *et al.*, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5200–5204.
[68] J. Wagner, D. Schiller, A. Seiderer, and E. André, "Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?" in *Proc. INTERSPEECH*, 2018, pp. 147–151.
[69] S. P. Dubagunta, B. Vlasenko, and M. Magimai-Doss, "Learning voice source related information for depression detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6525–6529.
[70] S. R. Kadiri and P. Alku, "Excitation features of speech for speaker-specific emotion detection," *IEEE Access*, vol. 8, pp. 60382–60391, 2020.
[71] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 469–472, Jun. 2009.
[72] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 614–624, May 2009.
[73] M. Airas, H. Pulakka, T. Bäckström, and P. Alku, "A toolkit for voice inverse filtering and parametrisation," in *Proc. INTERSPEECH*, 2005, pp. 2145–2148.
[74] T. Drugman, A.Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. INTERSPEECH*, Firenze, Italy, 2011.
[75] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," in *Proc. INTERSPEECH*, 2019, pp. 3920–3924.
[76] D. Tang, J. Zeng, and M. Li, "An end-to-end deep learning framework with speech emotion recognition of atypical individuals," in *Proc. INTERSPEECH*, 2018, pp. 162–166.
[77] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[78] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[79] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, pp. 206–215, 2019.
[80] C. A. Rosen, D. Anderson, and T. Murry, "Evaluating hoarseness: Keeping your patient's voice healthy," *Amer. Fam. Physician*, vol. 57, no. 11, pp. 2775–2782, 1998.

N. P. Narendra received the B.E. degree in electronics and communication Engineering from the Siddaganga Institute of Technology (affiliated to VTU), Tumkur, India, in 2009, and the M.S. and Ph.D. degrees from the Indian Institute of Technology Kharagpur (IIT-Kharagpur), Kharagpur, India, in 2012 and 2016, respectively. From January 2017 to August 2020, he is a Postdoctoral Researcher with the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. His research interests include speech synthesis, analysis and detection of pathological speech, and paralinguistic speech processing.

Björn Schuller, (Fellow, IEEE) received the Diploma in 1999, the doctoral degree for his study on Automatic Speech and Emotion Recognition in 2006, and his habilitation and Adjunct Teaching Professorship in the subject area of Signal Processing and Machine Intelligence in 2012, all in electrical engineering and information technology from TUM in Munich, Germany. He is Full Professor of Artificial Intelligence, and Head of GLAM - the Group on Language, Audio & Music, Imperial College London, U.K., Full Professor and ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany, Co-Founding CEO and current CSO of audEERING, and an Associate of the Swiss Center for Affective Sciences at the University of Geneva. Dr. Schuller is president emeritus of the Association for the Advancement of Affective Computing (AAAC), elected member of the IEEE Speech and Language Processing Technical Committee,, and senior member of the ACM. He coauthored five books and more than 700 publications in peer reviewed books, journals, and conference proceedings leading to more than 20 000 citations.

Paavo Alku (Fellow, IEEE) received the M.Sc., Lic.Tech., and Dr.Sc.(Tech.) degrees from the Helsinki University of Technology, Espoo, Finland, in 1986, 1988, and 1992, respectively. He was an Assistant Professor with the Asian Institute of Technology, Bangkok, Thailand, in 1993, and an Assistant Professor and Professor with the University of Turku, Turku, Finland, from 1994 to 1999. He is currently a Professor of speech communication technology with Aalto University, Espoo, Finland. He has authored or coauthored about 220 peer-reviewed journal articles and about 220 peer-reviewed conference papers. His research interests include analysis and parameterization of speech production, statistical parametric speech synthesis, spectral modeling of speech, speech-based biomarking of human health, and cerebral processing of speech. He is an Associate Editor for the *Journal of the Acoustical Society of America* He was an Academy Professor assigned by the Academy of Finland during 2015–2019. He is a Fellow of ISCA.